Data Catalogue

The CESSDA Data Catalogue contains the metadata of all data in the holdings of CESSDA’s service providers. It is a one-stop-shop for search and discovery, enabling effective access to European social science research data.

Data Management Expert Guide

This guide is designed by European experts to help social science researchers make their research data findable, accessible, interoperable and reusable.

Training

The CESSDA Training website provides a collection of resources and events for learning about the management, preservation and distribution of research data.
1. PLAN

- Personal data
- FAIR data principles
- Data management plan (DMP)
- DMP content elements
- Answer DMP questions and develop your own DMP
FAIR Data Action Plan
Interim recommendations and actions from the European Commission

Findable
To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accessible
Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

Interoperable
(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

Reusable
(Meta)data are sufficiently well described for both humans and computers to be able to understand them and have a clear and accessible data usage license.
Data Management Plan (DMP)

Adapt your DMP section at the end of every chapter

Corresponding questions to each chapter

<table>
<thead>
<tr>
<th>Title of the project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date and version of this plan</td>
</tr>
<tr>
<td>Description of the project</td>
</tr>
<tr>
<td>Origin of the data</td>
</tr>
<tr>
<td>Principal and collaborating researchers</td>
</tr>
<tr>
<td>Funder (if applicable)</td>
</tr>
<tr>
<td>Data producer</td>
</tr>
<tr>
<td>Project data contact</td>
</tr>
<tr>
<td>Data owner(s)</td>
</tr>
<tr>
<td>Roles</td>
</tr>
<tr>
<td>Costs</td>
</tr>
</tbody>
</table>
Adapt your Data Management Plan

A list of Data Management Questions based on the Expert Tour Guide on Data Management

---

**Overview**

- **Title of the project**
- **Date of this plan**
- **Description of the project**
  - What is the nature of the project?
  - What is the research question?
  - What is the project timeline?
- **Origin of Data**
  - What kind of data will be used during the project?
  - If you are reusing existing data: What is the scope, volume and format? How are different data sources integrated?
  - If you are collecting new data can you clarify why this is necessary?
- **Principal researchers**
  - Who are the main researchers involved?
  - What are their contact details?
- **Collaborating researchers (if applicable)**
  - What are their contact details and their roles in the project?
- **Funder (if applicable)**
  - If funding is granted, what is the reference number of the funding granted?
- **Data producer**
  - Which organisation has the administrative responsibility for the data?
- **Project data contact**
  - Who can be contacted about the project after it has finished?
- **Data owner(s)**
  - Which organisation(s) own(s) the data?
  - If several organisations are involved, which organisation owns what data?
- **Roles**
  - Who is responsible for updating the DMP and making sure that it's followed?
  - Do project participants have any specific roles?
  - What is the project timeline?
- **Costs**
  - Are there costs you need to consider to buy specific software or hardware?
  - Are there costs you need to consider for storage and backup?
  - Are potential expenses for (preparing the data for) archiving covered?
2. ORGANISE & DOCUMENT

- Organising data for research and data sharing
- Elements of data structure
- File naming, folder structure
- Data documentation
- Metadata standards

Organising data for research and data sharing

- Elements of data structure
- File naming, folder structure
- Data documentation
- Metadata standards
Data file structure

- Units of analysis / analytical objectives / methods of analysis
- Relations: different content items / sources of data / other relevant external information
- Connections to other existing or future data
- Strategies for version control
- Technical limitations (e.g. the size, software)
- Software

<table>
<thead>
<tr>
<th>CAPI Module</th>
<th>Name</th>
<th>All</th>
<th>Financial Respondent</th>
<th>Household</th>
<th>Family</th>
<th>non-proxy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>Coverscreen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DN</td>
<td>Demographics</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PH</td>
<td>Physical Health</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BR</td>
<td>Behavioural Risks</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CF</td>
<td>Cognitive Function</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH</td>
<td>Mental Health</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>Health Care</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EP</td>
<td>Employment and Pensions</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS</td>
<td>Grip Strength</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WS</td>
<td>Walking Speed</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CH</td>
<td>Children</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP</td>
<td>Social Support</td>
<td>(partly)</td>
<td></td>
<td></td>
<td>(partly)</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>Financial Transfers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HO</td>
<td>Housing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH</td>
<td>Household Income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>Consumption</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS</td>
<td>Assets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>Activities</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EX</td>
<td>Expectations</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>Interviewer Observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New modules in wave 2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td>Chair Stand</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF</td>
<td>Peak Flow</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XT</td>
<td>End-of-Life Interview</td>
<td></td>
<td>proxy interview, deceased</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dive in deeper - variable names and labels

File naming strategies

Folder structure

Flat (rectangular) data files
Hierarchical files
Relational database

Example:
20130311_interview2_audio.wav
20130311_interview2_trans.rtf
20130311_interview2_image.jpg
Documentation & metadata

- Two levels of documentation: (1) project level documentation; (2) data level
- Quantitative and qualitative sections on data level

Create a Codebook

A codebook is an essential document that informs a dataset's record layout, list of variables using the DDI metadata standard, in

Why a Codebook?
Creating a readable codebook to accurately represent a dataset, authoritative (straight-from-the-source).
To create a codebook, information about the dataset, including variables, values, and notes, is included.

Create machine-readable metadata

Check out The Dublin Core Metadata Generator (dublincoregenerator, n.d.) and see how metadata elements are converted into a machine-readable file in .xml.
Also, if you enjoy working with .xml schemas, get started in creating a codebook to accompany your dataset with the DDI codebook (DDI Alliance, 2017a).
3. PROCESS

- Data entry
- Data coding (quantitative, qualitative)
- File formats
- Data integrity and authenticity
- Systematic approach to data quality
Data entry and integrity

- Data integrity: assurance of the accuracy, consistency and completeness of original information in the data based on its structure and on links between data and integrated elements of documentation.

**Quantitative**
- Check the completeness of records
- Reduce burden at manual data entry
- Minimise the number of steps
- Conduct data entry twice
- Perform in-depth checks for selected records
- Perform logical and consistency checks
- Automate checks whenever possible

**Qualitative**
- Prevent mistranscription by recording high-quality data
- Determine the transcription method
- Choose between manually transcribing or with the help of speech recognition software (SRS)
- Determine the rules
- Transcribe
- Check the transcription
- Protect your participants
- Choose a QDA-compatible file format
- Choose a file format for long-term preservation
Quantitative coding/qualitative coding

Quantitative:
- General rules, recommendations/check lists
- Documentation: subsection - organising variables (integrated doc./internal structure of the data file)
- Standardised coding schemes
- Missing values
- Coding variance

Qualitative:
- Coding is a way of indexing or categorizing the text in order to establish a framework of thematic ideas about it | Gibbs (2007)
- Concept driven coding versus data driven coding
Dive in deep? Weights of survey data

Adjustment of the sample. Each individual case in the file is assigned an individual weight which is used to multiply the case in order to attain the desired characteristics of the sample.

There are different types of weights for different purposes

- Necessary in some situations
- Issue of quality

Distribution of weights

If the weight of a case equals 1 then the values measured are not adjusted. In the case of post-stratification weights both high or low numbers indicate either large deviations of the sample from the target population, poor quality of the weight or both. It is desirable the large part of values of the weighting variable is close to 1.

Weights constructed by others

Is there any weighting variable in your working data file? If yes and you are not the author of the weight, never use it without knowledge of its origin and purpose. You should always thoroughly explore the distribution of the weighting variable and its impact on distributions of other selected variables from the data file.
### File formats and data conversion

- **Short-term data processing:** file formats for operability
  - Proprietary vs. open formats
  - Export / portable formats
- **Long-term data preservation**
- **Link to the table of Recommended file formats**

### File Formats and Data Conversion

<table>
<thead>
<tr>
<th>Type of data</th>
<th>Recommended formats</th>
<th>Acceptable formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabular data with extensive metadata</td>
<td>SPSS portable format (.por) delimited text and command (setup) file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file</td>
<td>proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)</td>
</tr>
<tr>
<td>Tabular data with minimal metadata</td>
<td>comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements</td>
<td>delimited text (.txt) with characters not present in data used as delimiters</td>
</tr>
<tr>
<td>Geospatial data vector and raster data</td>
<td>ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) geo-referenced TIFF (.tif, .tiff) CAD data (.dwg) tabular GIS attribute data Geography Markup Language (.gml)</td>
<td>ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mil) for vector data Keyhole Mark-up Language (.kml) Adobe Illustrator (.ai), CAD data (.dxf or .svg) binary formats of GIS and CAD packages</td>
</tr>
<tr>
<td>Textual data</td>
<td>Rich Text Format (.rtf) plain text, ASCII (.txt) Extensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema</td>
<td>Hypertext Mark-up Language (.html)</td>
</tr>
<tr>
<td>Image data</td>
<td>TIFF 6.0 uncompressed (.tif)</td>
<td></td>
</tr>
</tbody>
</table>

---

[Link to the table of Recommended file formats]
Data authenticity & version control

Best practices for quality assurance, version control and authenticity

Version and edition management will help to:

1. Clearly distinguish between individual versions and editions and keep track of their differences;
2. Prevent unauthorised modification of files and loss of information, thereby preserving data authenticity.

Best practices
The best practice rules (UK Data Service, 2017a; Krejčí, 2014) may be summarised as follows:

- Establish the terms and conditions of data use and make them known to team members and other users;
- Create a ‘master file’ and take measures to preserve its authenticity, i.e. place it in an adequate location and define access rights and responsibilities – who is authorised to make what kind of changes;
- Distinguish between versions shared by researchers and working versions of individuals;
- Decide how many versions of a file to keep, which versions to keep (keep version 02-00 but not 02-01), for how long and introduce clear and systematic naming of data file versions and change records;
- Record relationships between items where needed, for example against, between data file and related documentation or metadata;
- Document which changes were made in any version;
- Keep original versions of data files, or keep documentation that can identify them;
- Track the location of files if they are stored in a variety of locations;
- Regularly synchronise files in different locations, such as using cloud storage;

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Created By</th>
<th>Date Created</th>
<th>Maintained By</th>
<th>Version Number</th>
<th>Modified By</th>
<th>Modifications Made</th>
<th>Date Modified</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Wrap up: Data quality

- **Small things matter:** "The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to [preventing, measuring and] dealing with the many important problems that can arise." American Association for Public Opinion Research (2015) (AAPOR)

- **In qualitative research,** discussions about quality in research are not so much based on the idea of standardization and control, as this seems incompatible with many qualitative methods. Quality is rather seen as an issue of how to manage it. Sometimes it is linked to rigour in applying a certain method, but more often to soundness of the research as a whole" | Flick (2007).

- **A complex approach to data quality:** "The mechanical quality control of survey operations such as coding and keying does not easily lend itself to continuous improvement. Rather, it must be complemented with feedback and learning where the survey workers themselves are part of an improvement process" Biemer & Lyberg (2003).
4. STORE

- Storage solutions
- Storage strategies
- Disaster recovery strategies
- Protect: passwords and encryption
Towards a Storage Strategy

- A Storage strategy contains:
  - storage solutions and media
  - backup strategy and disaster recovery
  - data protection
- systematically implemented in a data management plan
7. Discover

This upcoming chapter is for data users, i.e. people who are looking for research data. It will be available toward the end of 2018.

Main take-aways - after reading through this chapter you should:

- Be aware of different types of data resources for social sciences
- Know more about ways of searching for social science data
- Be able to use search engines in data repositories effectively
- Be aware of steps in evaluating the quality and usefulness of data for secondary analysis
- Understand different types and modes of access to data
- Be informed on research data relevant for selected research topics and recommended by experts.
Thank you
The Data Management Expert Guide has been created for CESSDA ERIC by a number of its service providers' experts at: ADP, AUSSDA, CSDA, DANS, FORS, FSD, GESIS, NSD, SND, So.Da.Net and UKDS and is illustrated and edited by Verbeeldingskr8.